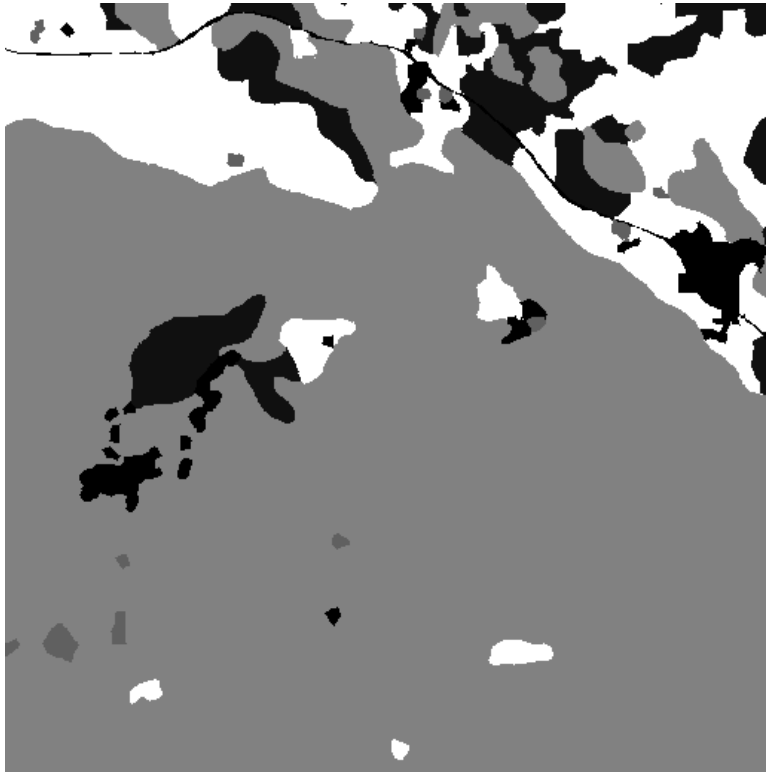
Automatic Cataloguing and Characterization of Earth Science Data Using SE-trees

Ron Rymon
Intelligent Systems Program,
University of Pittsburgh
Telephone: (412) 624-2287
E-mail: Rymon@ISP.Pitt.edu
WWW: http://www.isp.pitt.edu/~rymon

February 1996

Overview of NASA Task and Approach

- EOS Satellites will produce enormous amounts of remote sensing data.
- Need: content-based storage/access and analysis tools.
- Challenge: a universal classifier that will recognize ground truth from sensory input.
- Potential advantages of our approach:
  - Accuracy;
  - Noise tolerance;
  - Flexible tradeoff between time/space and accuracy;
  - Use of other knowledge sources.

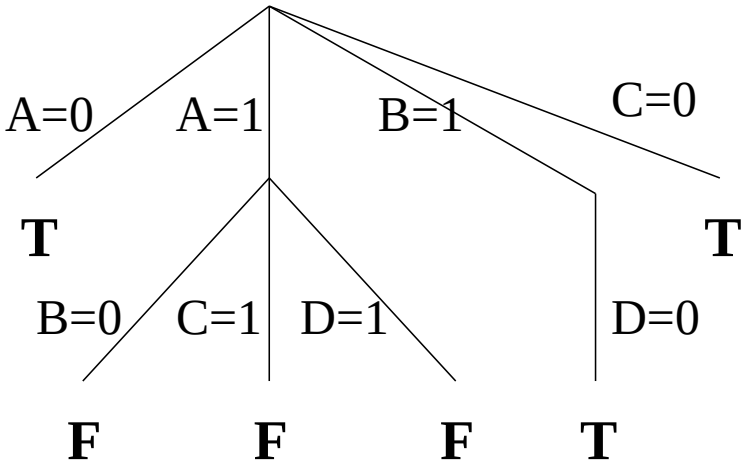


Approach: SE-tree-based Induction

Generalizes Decision Trees

- Learning: recursive partitioning on *multiple* attributes

A	B	C	D	Class
0	0	1	0	T
0	1	1	1	T
1	0	1	0	F
1	1	0	0	T
1	1	1	1	F



A=0 ⇒ T
A=1∧B=0 ⇒ F
A=1∧C=1 ⇒ F
A=1∧D=1 ⇒ F
B=1∧D=0 ⇒ T
C=0 ⇒ T

- Classification: traverse matching paths, e.g. {A=1,B=0,C=0,D=1}

- New "difficulties":
  - Combinatorics ⇒ construct partially, using **exploration policy** (bias).
  - Inconsistency ⇒ use **resolution criteria** (bias).

Advantages of SE-tree-Based Induction

- Tree/Rules are symbolic, interpretable by humans:

  $(58<\text{Band1}<76)(132<\text{Band3}<145)(\text{Time}=\text{Afternoon})\Rightarrow\text{Urban}$

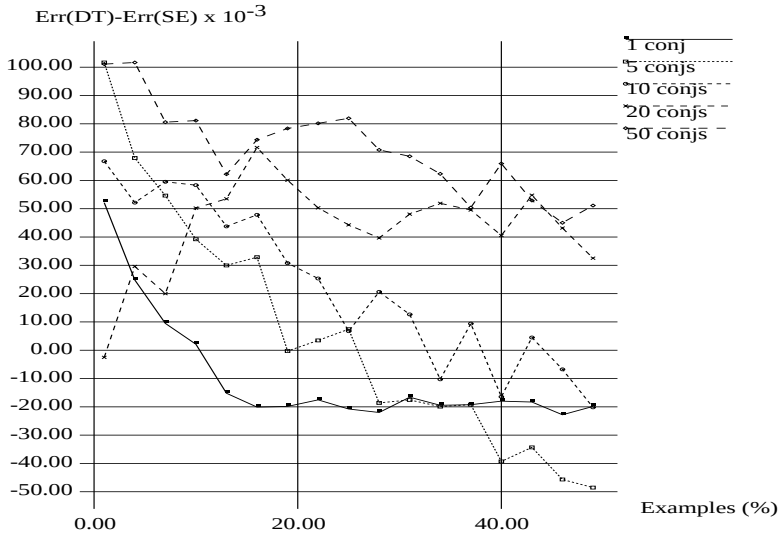
- Hypothesis-space bias can be introduced to reflect domain:
  - Resolution criterion;
  - Exploration policy;
  - Overruling theory/constraints.

- Spectrum of tradeoff between accuracy and size/space:
  - Exploration until diminishing return, resource limitation.

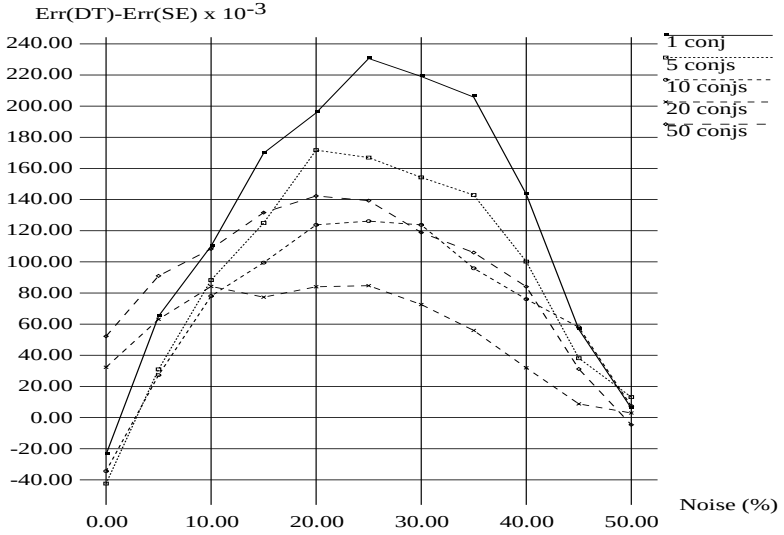
- Can combine induced knowledge with other knowledge sources.

Advantages over Decision Trees

- Can extract more information from fewer examples:



- Less sensitive to noise:



SE-Learn software package

General Features:

- Modes: Learn, Learn+Test, Test, Produce, X-validatation, Seq-validation;
- Attribute types: Nominal, Integer, Continuous;
- Automatic/manual discretization of ordered variables;
- Attribute selection: Entropy (ID3, C4.5), GINI (CART), χ^2 (ChAID);
- Exploration policies: Primary DT, Cardinality, Entropy, Beam, Best;
- Significance testing: Binomial, Chi-square, Entropy threshold;
- Resolution criteria: voting, prefer more general/specific rules.

Implementations:

- LISP prototype;
- Commercially developed C version;
- SE-image.

The Image Interpretation Problem and Goals

- Given:
  - Sensory inputs (radiances, per pixel);
  - Other pixel-based data, e.g. elevation, texture;
  - Other general data, e.g. season, weather.
- Classify each pixel, e.g. Urban, Agricultural, etc.

Research Goals:

- Improved accuracy and noise resiliance;
- Improved and smoother size/accuracy tradeoff;
- Use of domain-specific knowledge (mostly in bias formation).

Development Goals:

- Making SE-Learn/SE-image available to domain scientists.
- Linking SE-Learn to current image interpretation tools;

Research Results

- Following experimental design of (Chettri *et al.*, 92)

	Accuracy
Back Propagation Neural Network	72.7%
Gaussian Maximum Likelihood Classifier	65.3%

- *SE-Learn*

	Accuracy	Size
As is	80.0%	
With 4 neighbors	89.8%	O(1M)
+Discretization		
Pure clusters	83.2%	
Bit-discretized	81.6%	O(100K)
+Statistical pruning		
Bit-discretized	78.5%	O(10K)
+Bottom-up merging discretization	80.1%	202 rules

Research Results (cont.)

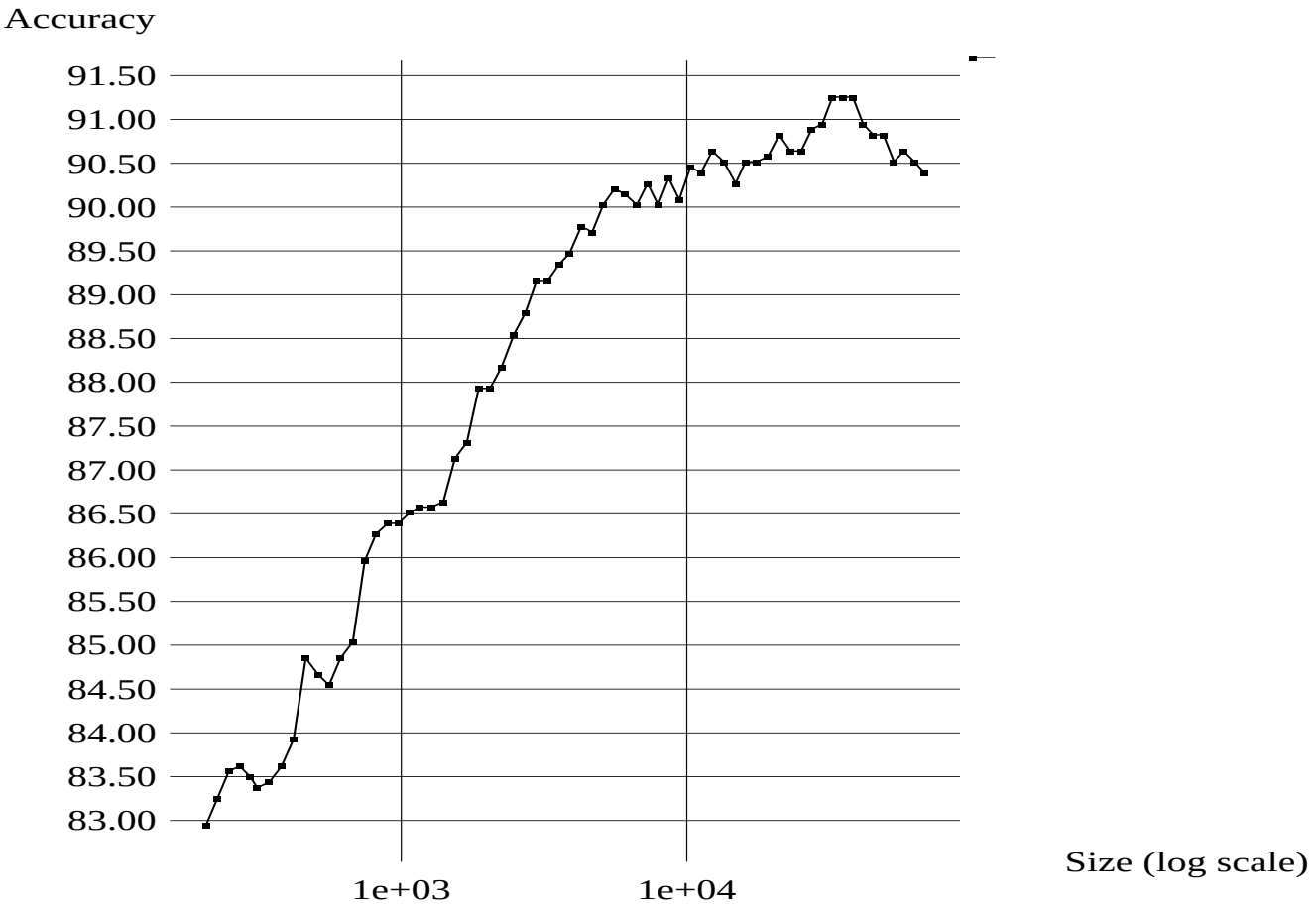
Discriminating forest ecosystems types:

- Work with Civco, Silander (U Conn), Wang (U Ill) & Gong (UC Berkeley).
- Input: 6 TM bands over NW CT, for May, Aug, and October; output of an illumination model, and a Road/No-Road discriminator.
- Accuracy: 91.3% on unseen data:

	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18	19	20	21	22	23	Total
1	87							1				1	1	5			2							97
2	2	34						3									6							45
3			78	4				1					3							1				87
4			2	79	1			3						1	2									88
5				3	93			1									7							104
6						80	2						1								1			84
7						2	76	3			11						1				2			95
8								192				4						1		7				204
9								3	164			6					1		6	1				181
10										22														22
11							9	3			80									1				93
12												6												6
13										1		3	56	1										61
14				1										42			3							46
15													1		46									47
16								2								25			1					28
17								1			1						36			1				39
18						1		2				1						19						23
19								6	1			2				1			76					86
20																								0
21								2	1							2					52		1	58
22																						80		80
23						1		1											1		2		45	50
Total	89	34	80	87	94	84	87	224	166	23	92	23	62	49	48	28	56	20	84	11	57	80	46	1624

Research Results (cont.)

- Size/accuracy tradeoff



Development Results

- Completed LISP prototype of SE-Learn
  - ported to a NASA machine (short@danville.gsfc.nasa.gov).
  - freely available to anyone.

- C version developed commercially by Modeling Labs
  - freely available to scientists.

- SE-image: specialized version for image classification
  - applicable to any image (binary format)
  - allows (rather limited) manipulation of SE-Learn's parameters

Summary

- SE-tree-based induction:
  - Symbolic, human interpretable models;
  - Fairly resiliant to noise;
  - Wide range of tradeoff between time/space and accuracy;
  - Can utilize domain knowledge;
  - SE-Learn/SE-image implementations freely available.

- Interested in embedding SE-Learn into other software packages

- Interested in collaborations with domain experts on specific research projects.